ALGORITMO SIN ETIQUETAS, NO SUPERVISADO

**¿Qué son las técnicas de agrupamiento?**

Técnica de Machine Learning que consiste en dividir población en grupos mas similares (por la distancia o Angulo).

Es casi imposible diseñar una estrategia por cada individuo,

Pero si agruparlos por similitudes a los clientes

Existen dos tipos de agrupamiento:

• Agrupamiento estricto (hard clustering): cada dato pertenece a un grupo u otro, no hay puntos medios.

• Agrupamiento laxo (soft clustering): en lugar de asignar un dato a un grupo, se asignan probabilidades a cada dato de pertenecer o no a un grupo.

Un punto muy importante que debes considerar cuando ejecutas técnicas de agrupamiento es que debes definir muy claro a qué te refieres cuando hablas de similitud entre puntos, porque esto puede ayudarte a definir el algoritmo correcto para tus necesidades particulares.

Modelos para determinar similitudes

A grandes rasgos, existen cuatro aproximaciones para definir similitud:

• Modelos conectivos: los puntos más similares son los más cercanos en el espacio de búsqueda. Recuerda que este espacio puede ser altamente dimensional cuando tus feature vectors definen muchas características a analizar. Una desventaja de este tipo de modelos es que no escalan para conjuntos de datos grandes, aunque es posible utilizar una muestra y aplicar técnicas de estadística inferencial para obtener resultados.

•

• Modelos de centroide: definen similitud en términos de cercanía con el centroide del grupo. Los datos se agrupan al determinar cuál es el centroide más cercano.

•

• Modelos de distribución: trata de asignar probabilidades a cada dato para determinar si pertenecen a una distribución específica o no (por

ejemplo, normal, binomial, Poisson, etc.).

•

• Modelos de densidad: analizan la densidad de los datos en diferentes regiones y dividen el conjunto en grupos, luego asignan los puntos de acuerdo a las áreas de densidad en las que se haya dividido el dataset.

Utiliza varios modelos con el mismo conjunto de datos para analizar el rendimiento de los diversos algoritmos

Agrupamiento jerarquico

Agrupa objetos similares (cluster)

Inicia tratando a cada objecto como un cluster individual y luego recursivamente

- Identifica los 2 cluster con menor distancia (mas similares)

- Agrupa los 2 clusters en uno nuevo

El output final es un dendograma que muetra relacion entre objetos y grupos

Importante determinar que medida de distancia usar y los puntos a utiiizar en cada cluster

Chart

Description automatically generated

K-means

Agrupa con centroides

Asigna puntos al azar (K define el numero inicial de clusters)

En cada iteracion el punto se ajusta a su nuevo centroide

y cada punto se recalcula con la distancia con respecto de los centroides

Los puntos se reasignan al nuevo centro

El algoritmo se repite iterativamente hasta que no hay mas mejoras